学术论文作者同名消歧方法研究进展

王 新，卢 垚，袁 雪，赵婉婧，陈 莉，刘敏娟*

（中国农业科学院农业信息研究所，北京 100081）

摘 要：[目的／意义]调研近年来作者同名消歧相关研究，厘清发展脉络，为后续研究提供参考。[方法／过程]使用 Web of Science、Scopus、谷歌学术、ACM、IEEE、Elsevier、Springer、中国知网、维普数据库和万方数据库检索作者姓名消歧相关文献，选择其中 46 篇代表性文献进行综述。从数据对作者同名消歧方法的影响的角度审视、梳理相关研究的发展脉络。[结果／结论]按照消歧任务所依据的数据特点将相关研究方法分为 3 类。随着技术的进步，深度学习方法得到广泛采用。相对于模型的改进，基于深度学习的特征学习和表示，对作者同名消歧算法效果的提高更为显著，同时，为充分利用数据中包含的各种信息，3 类算法呈现出相互结合、互补增益的态势。从文献调研情况看，可以从增量消歧和跨语种消歧等角度开展后续研究。

关键词：知识组织；作者名消歧；人名消歧

中图分类号：G353.1 **文献标识码：**A **文章编号：**1002-1248（2022）10-0082-09

引用本文：王新，卢垚，袁雪，等. 学术论文作者同名消歧方法研究进展[J]. 农业图书情报学报，2022，34(10)：82-90.

1 引 言

随着信息技术和出版行业的快速发展，科研产出数量涨势迅猛。据统计，全球约有 10 000 家期刊出版商，每年发表 300 多万篇文章，近年来论文年增长率为 4%，期刊数量年增长率已经超过 5%[1]。然而，在如此庞大的文献数量之下，由作者同名现象所引起的作者实体指代不明的问题频频出现，为科研成果管理、科学文献管理、文献搜索服务、社交网络分析等诸多应用场景带来了挑战。该问题作为学术评价、科学家流动等人文社科类研究课题的基础性重要环节，严重掣肘相关研究的开展，成为亟待解决的问题之一。

学术论文作者同名消歧任务所面临的主要情况是同名异人，即多名作者共享同一个姓名。作者同名消歧任务的目标是建立论文与现实中作者实体的关系，问题的核心是判断出现在多篇论文中的同一作者姓名是否指向现实生活中的同一个作者个体，难点是同名异人的情况往往涉及同一领域，甚至同一机构的同名作者区分。

本文使用 Web of Science（选择库 WOS 核心集、CABI、中国科学引文数据库、SciELO Citation Index）、

收稿日期：2021-11-22

基金项目：中国农业科学院农业信息研究所 2022 年科技创新工程“数字农科院 3.0 建设”（CAAS-ASTIP-2016-AII）

作者简介：王新，馆员，研究方向为数字资源管理、信息组织。卢垚，副研究员，研究方向为农业信息资源建设与组织。袁雪，馆员，研究方向为信息资源组织与管理。赵婉婧，助理研究员，研究方向为信息组织技术与实践。陈莉，馆员，研究方向为信息资源组织

*通信作者：刘敏娟，副研究馆员，研究方向为信息资源组织与管理。E-mail：liuminjuan@caas.cn

Scopus、谷歌学术、ACM、IEEE、Elsevier、Springer、中国知网、维普数据库和万方数据库，检索作者姓名消歧相关文献，检索词选用英文“name disambiguat*”“author disambiguat*” “disambiguation of names”和中文“姓名消歧” “作者消歧” “著者消歧”，采用“论文作者姓名消歧” “Name disambiguat* for author”“author name disambiguat*”等限定词进行精炼，将研究方向限定为计算机科学和情报学图书馆学，检索结果经汇总、排除重复、人工剔除误检、根据参考文献补充漏检后获得文献 470 余篇，涵盖了从 1998 年至 2021 年的相关文献，限于篇幅，在兼顾权威性、影响力和新颖性的前提下，选择其中 46 篇文献进行归纳总结。

作者同名消歧最早是由 BAGGA 和 BALDWIN[2]于 1998 年首次提出，此后逐步引起学界关注，并于 2001 年举办的数字图书馆联合会议（JCDL）上作为主要议题进行讨论。经过 20 多年的不断探索，学界对此问题进行了大量研究，也取得了相当丰硕的成果。通过对近 3 年来的综述类文献的研读，发现既往综述文献有的聚焦于单一文献数据库的方法述评[3]；有的针对特定算法进行分析，未能反映问题全貌[4]；也有文献着重对网页人名消歧研究的整理，场景偏于网络[5]；更有学者对相关技术方法从不同视角做了分类整理和汇总，但并未关注数据对问题解决方法的影响，也未阐明数据与算法之间的关系[6]。为此，本文以机器学习视角，从消歧任务所采用特征数据的结构入手，将相关研究划分为基于文献特征的消歧方法、基于社会网络的消歧方法和整合外部知识的消歧方法 3 个类别，从数据层面审视数据对作者同名消歧方法的影响，厘清发展脉络，为后续研究及应用提供参考。

2 数据分析及存在问题

2.1 数据分析

作者同名消歧问题属于命名实体消歧的范畴[7]，可以归为分类或聚类问题，处理过程一般包括数据收集和分析、数据预处理、特征抽取、分类训练或聚类、结果验证等步骤。由于特征来源于对数据的学习，影响机器学习效果的主要决定因素在于数据。数据可以有多种类型和结构，例如文献特征信息一般以非结构化文本呈现，提取后的特征可以用二维表来存储和表示；引证信息和人际关系是网状的关系型数据，可以用图、键值对或二维表来存储、表示其二元或多元关系。数据结构不同，根本原因在于其语义的差异，但数据结构本身决定了其适用算法。作者同名消歧任务的数据来源可以非常直观地分为 3 个部分：①文献特征。文献特征包括标题、关键词、摘要、全文、出版信息、引文信息等。②作者特征，包括作者的姓名、邮箱、隶属机构、地址及其他信息。从作者信息中可以提取作者之间的合著关系，构成合著网络。③外部引入特征，即除去论文所提供的信息以外，通过作者个人主页、百科等其他外部信息来源获取的特征信息。

无论是文献特征，还是作者特征，大部分特征信息所描述的都是一种二元关系，可以用二维表来表示和处理，但是引文信息和作者特征中的合著关系描述的是多元关系，适用键值对或图结构来存储和处理。

2.2 数据中存在的问题

在作者同名消歧任务中，数据方面存在的问题为同名消歧问题的解决带来了困难和挑战，主要有以下 3 个方面。

（1）作者信息不完善。学术论文元数据中关于作者的信息不够完善。从论文提交、发表，到元数据采集汇总，面向科研管理或学术评价等研究提出消歧任务，中间的各个环节都会影响数据的完整性。在论文提交发表环节，不同出版机构对著录信息的完备性要求存在差异，著录要求不尽相同，作者的属性信息丰富程度参差不齐；元数据采集整理环节，数据采集往往出现数据缺项、漏采，数据清洗规范过程不可避免对数据造成二次破坏。学术论文作者消歧任务所使用的作者数据，其信息往往不完整，缺乏足够的信息来作为消歧的依据。

（2）作者信息不规范。作者信息不规范表现为同一著者存在多个名字。这一现象主要出现在外文期刊

论文中，造成这种情况的原因主要是外文姓名拼写存在多种变体（如全称和缩写），同时也存在元数据录入错误的情况，造成姓名著录格式不规范的结果。

（3）作者信息动态变化。作者在现实中的所属单位、地点、联系方式等信息存在变更的可能，同一作者不同时期的论文，其属性信息存在前后不一致的情况。

3 作者同名消歧的研究方法

根据消歧任务所使用特征数据结构的不同，该问题的研究方法也大致可以划分为 3 个类别：基于文献特征的消歧方法、基于社会网络的消歧方法和整合外部知识的消歧方法。其中，基于文献特征的消歧方法，根据特征处理方式的不同，还可以分为有监督的消歧方法、无监督的消歧方法和半监督的消歧方法。有监督的消歧方法将作者同名消歧视为分类任务，采用事先已标注的数据训练模型，再应用模型对未区分同名作者的方式进行分类，优点是精度较高，缺点是所需的训练集数量庞大且获得成本高。无监督的消歧方法将同名消歧视为聚类任务，只要输入模型的特征具备足够的区分度，就可以从数据中自动习得模式且获得相当不错的效果，优点是不依赖训练数据，缺点是在大数据集上收敛较慢，对参数比较敏感。

3.1 基于文献特征的消歧方法

基于文献特征的消歧方法是同名消歧命题提出之初所采用的研究方法，此方法的主要思路是从论文元数据中提取各种特征，即利用论文的标题、关键词、摘要、作者单位等文献外部特征，寻找或构造对作者姓名具有最大区分度的特征集合，然后进一步优化、提取、保留有效特征，删除无关特征，一方面实现数据降维，降低噪声，减少计算的资源开销；另一方面提高模型的鲁棒性和泛化能力，最后选择合适的算法计算得出消歧结果。

算法选择方面，有监督方法常用算法是空间向量模型（Vector Space Model），无监督方法采用的算法是各种聚类算法，二者经常配合使用，目的是提高分类或聚类的准确率。为达到这一目的，相关研究从两个方向进行了探索，一种是利用空间向量模型进行特征提取和表示，然后作为聚类算法的输入数据实现作者同名消歧。例如提出同名消歧问题的 BAGGA 和 BALDWIN 利用向量空间模型算法对跨文档的共指链信息进行向量化表示，通过评分实现跨文档共指消歧[2]。CHENG 等[8]抽取全文中的命名实体作为特征，包括名词短语和人物，并使用 Soft-TFIDF 算法将其表示为特征向量，通过计算特征向量之间的相似度作为层次聚类方法的输入值。章顺瑞等[9]抽取文章中的命名实体、名词短语和动词短语作为特征词，利用空间向量模型处理为特征向量后使用词频加权，作为层次聚类算法的输入以提高聚类效果。另一种是分阶段多次聚类，不断精练聚类结果从而提高精度。例如丁海波等[10]提出一种三阶段聚类的方法，通过在第二、第三阶段聚类过程中加入上下文特征以解决属性数据的稀疏性问题，提高系统召回率。WANG 等[11]提出了基于两步策略的自适应共振理论（ART），分两步聚类模拟人工消歧过程。

在对特征的选择方面，相关研究经历了人工抽选特征到自动学习语义特征的过程。早期研究关于特征选择，一般以人工选择为主，例如 LONG 等[12]抽取命名实体和名词作为特征词，同时利用特征词与作者姓名的句间距作为权值对特征词加权，构成特征向量。随着时间发展和相关研究的推进，特征选择也呈现出自动化、抽象化的特点，如 ANDERSON 等[13]提出一种 SAND 自学习器，通过启发式方法自动学习作者姓名、作品和地点名称作为训练数据的特征，解决了特征稀疏问题。PEDERSEN 等[14]利用奇异值分解（SVD）对文本特征向量进行降维，提高了相似度计算的准确率。伴随着深度学习技术的兴起，该技术也被引入进来，作为自动学习语义特征的工具，如 TRAN 等[15]提出一种使用深度神经网络自动学习特征以解决作者姓名歧义。阮光册等[16]提出一种融合文献外部基本特征和内部语义特征的方法，利用 BERT 模型和词嵌入对文本内容的语义信息进行学习和向量化表示，最终将融合多特征的数据输入 XGBoost 完成作者同名消歧。马莹

莹等[17]提出基于特征编码和图嵌入的作者同名消歧方法，利用 word2vec 编码文档的属性特征，然后采用图自动编码器将文档关系编码到文档向量中，最后聚类实现作者同名消歧。

基于文献特征的消歧方法以文献的外在特征属性或内在语义特征作为消歧的依据，因此存在对数据使用不充分的短板，特别是对各种关系型数据的表示能力比较差，如作者合著关系、作者—所属机构关系等。

3.2 基于社会网络的消歧方法

基于社会网络的消歧方法将论文外在特征中的作者社会关系网络作为区分作者实体的依据，该方法利用作者在社会关系网络中的关系特征和社会关系的传递性构建社会网络图，根据合著、引用、隶属等关系构建边，以关系出现次数作为边的权重，然后利用图相关理论计算同名作者之间的相似度（拓扑距离），从而实现作者同名消歧。MALINA[18]根据一个实体名称可以在多个来源中出现的事实，首次提出构建基于作者姓名的关系网络，并从本地局部聚类和全局随机游走聚类两个角度分别给出了对应的解决方案。郎君等[19]利用检索结果中同名作者的共现现象发现并拓展其潜在社会网络，结合图的谱分割算法和模块度指标进行聚类，实现作者同名消歧。YAO[20]将姓名消歧方法应用于保险领域，先用属性匹配合并客户姓名，再利用链接分析客户网络结构，合并相同信息后实现人名消歧。NADIMI 等[21]结合启发式层次聚类和社交网络，将文献引用关系、作者关系等信息构建为社会网络图，实现人名消歧。

算法的演变主要表现在特征表示方面，一方面，特征表示由高维稀疏向量特征表示方式转变为经过深度学习处理的低维稠密特征向量表示方式。传统的社会网络采用维数较多的稀疏向量对特征进行表示，这种方法仅适用于小数据集低维求解，但是应用到大数据集时就会遭遇“维度爆炸”。为解决这一问题，基于社会网络的消歧方法借鉴了深度学习词向量的构建方式，将节点序列视作词序列，将高维稀疏向量映射为低维稠密向量的特征表示，显著改善了此种方法在大规模数据集上的应用效果。例如 PEROZZI 等[22]首次将语言建模思想引入社会网络的特征表示中，提出 DeepWalk 算法，将社会网络节点在连续向量空间中进行编码，以便统计模型利用。MIKOLOV 等[23]将 Skip-gram 模型引入社会网络方法，提出一个分层 softmax 替代方案通过对高频词进行二次采样和降采样特征学习，获得了显著加速。GROVER 等[24]提出 Node2Vec 算法作为对 DeepWalk 算法的改进，该方法采样随机游走实现更加全面的节点关系采样。另一方面，特征表示由单纯社会关系网络等同构网络向融合关系类型、节点属性等信息的异构网络方向演变。例如陈莉等[25]在 DeepWalk 算法的基础上进行改进，将节点之间边的关系类型考虑在内，提出 NEES 算法，能够采样得到边关系类型信息的边向量，同时能为图中每个节点学习到一个低维表示。WANG 等[26]提出一种结构化深度网络嵌入方法，即 SDNE 算法，用半监督的深度模型来捕捉高度非线性结构，通过结合一阶相似性（监督）和二阶相似性（非监督）来保留局部和全局特征。刘正铭等[27]提出一种融合节点文本属性信息的网络表示学习算法，建立基于参数共享的共耦神经网络，利用负采样和随机梯度下降优化策略实现模型快速收敛，从而获得融合网络结构信息和节点文本属性信息的特征表示。

基于社会网络的消歧方法侧重于对作者社会关系网络的数据表示，但是缺乏对文献外在特征的数据表示能力，因此也存在数据利用不充分的情况。

3.3 整合外部知识的消歧方法

鉴于论文数据中可用作者信息有限，研究者尝试通过整合外部资源和知识进而达到数据增强的效果。此类方法利用网络公开资源构建新的规则和类别，选取现实中人物信息中具有较强区分度且具备较高准确度的社会属性，建立其与待消歧姓名的联系，从而实现丰富人物特征的目的。

基于整合外部知识的消歧方法也经历了从传统实体链接方法走向结合深度学习方法获取特征抽象和泛化能力的过程。杨欣欣等[28]抽取网页文本中人名实体相关的依存特征和命名实体等辅助特征，采用二层聚

类实现人名消歧。HAN 等[29]提出一种利用专业类别知识从 Freebase 中自动挖掘参考实体表的 Web 查询方法，通过分类将人名链接到个人实体来实现消歧。VU 等[30]利用 Web 目录作为知识库，查找姓名在公共文档中的上下文，结合文档相似性实现人名消歧。SHEN 等[31]提出一种利用嵌入在维基百科中的丰富语义知识和知识库的分类法，将文本中的命名实体与统一维基百科和 WordNet 的知识库联系起来的方法。宁博等[32]从中文维基百科等知识库抽取人物信息、实体关系等实体信息对象，提出基于异构知识库的层次聚类方法。HAN 等[33]将知识库中的每条实体定义作为文本，从中抽取关于人物属性的 19 个特征形成向量，并以此辅助消歧。PENG 等[34]提出用于中文命名实体识别和消歧任务的 SIR-NERD 系统，该系统使用两阶段方法，先将知识库中同一人名下所有实体作为一篇文本并对其实体指称项分类，再对真正指向实体的指称项进行聚类以实现消歧。HE 等[35]利用深度神经网络堆叠降噪自动编码器学习初始文档的特征表示，然后通过微调优化相似性度量的表示，该方法在没有任何手动设计特征的情况下，在两个公共数据集上性能击败了复杂的集体方法。SUN 等[36]用卷积神经网络将实体表述及其上下文在连续向量空间中进行编码，并嵌入上下文词的位置以考虑上下文词和提及之间的距离，同时使用神经张量网络来模拟上下文和提及之间的语义交互，显著提高了消歧性能。FRANCIS 等[37]使用卷积神经网络来学习文本的上下文和实体的规范描述页面，提高了链接性能。CHEN 等[38]基于预训练的 BERT 模型将潜在实体类型信息注入实体嵌入中，另外，把基于 BERT 的实体相似度分数集成到最先进模型的本地上下文模型中，以更好地捕获潜在实体类型信息。

此外，基于整合外部知识的消歧方法与基于社会网络的方法相结合，从“局部”走向“全局”。GUPTA 等[39]探索大量维基百科的链接，使用多种信息源对实体描述、实体上下文及结构化知识学习统一的密集表示，无需特定领域训练数据或人工设计，解决了训练数据不足的问题。LE 等[40]将实体链接的文本共同引用关系作为潜在变量加入全局模型，以端到端的方式优化实体链接系统，取得良好性能。GUO 等[41]以语义相似的自然概念为指导，以迭代和贪婪近似方法，对文档中提到的所有实体消歧，利用在知识库产生的子图上进行随机游走获得的概率分布和词汇统计特征提高链接性能。YANG 等[42]提出动态上下文增强（DCA），顺序累积上下文信息以进行高效集体推理，并且可以作为插件和增强模块来应对不同的本地实体链接模型。针对实体链接系统对专门标注大量文档的依赖问题，LE 等[43]提出一种仅利用自然发生信息（未标记文档和维基百科）的方法，首先建立未标记文档中提及候选实体的高召回列表，然后使用候选列表作为弱监督来约束文档级实体链接模型。LE 等[44]还探索了在没有任何标记示例、只有知识库和来自相应领域的未注释文本集合的情况下学习链接引用的方法，将实体链接任务定义为多实例学习问题，依赖表面匹配创建初始化标签，将实体链接问题构造为远程学习问题。

整合外部知识的消歧方法本质是对消歧数据的补充和增强，并未从本质上解决前述两类方法数据利用不充分的问题。

4 结　语

纵观作者同名消歧研究的进展，不难发现，所有的消歧模型和算法都是围绕着数据的特点和不足展开的。可以说，有什么样的数据，就有什么样的算法，算法的提出与改进，都是针对数据的不足加以补充或妥协的结果。针对论文数据中作者属性信息不完善的情况，研究者另辟蹊径提出了整合外部知识的作者同名消歧方法；针对论文自身属性中内在特征和外在特征的不同特点，分别发展出了基于特征的作者同名消歧方法和基于社会网络的同名消歧方法；针对越来越大的数据体量和维度爆炸的特征模型，引入基于深度学习的词嵌入成为最佳方案。随着技术的进步，在特征选择方面，3 类研究方法逐渐发展为采用深度学习技术的特征表示方式，相对于模型的改进，基于深度学习的特征学习和表示，对作者同名消歧算法效果的提高更为显著。此外，上述 3 类方法也并非泾渭分明，

呈现出相互结合使用的态势，目的是尽可能多地利用各类信息，取长补短，以期获得更好的效果。例如，吴柯烨等[45]构建基于异源数据的二阶段姓名消歧框架，在充分挖掘本地关联信息的基础上，结合外源数据和本地关系发现实现全面姓名消歧；王若琳等[46]提取文本特征和文章与合著者之间的关系信息，采用论文嵌入网络构建异质信息网络，融合内容信息和关系信息，基于循环神经网络和层次聚类实现作者姓名消歧；郭晨亮等[47]使用词嵌入处理文献特征信息，结合元路径随机游走构建异构网络，最后以密度聚类算法完成消歧。

展望未来，作者同名消歧相关研究可以从以下两方面入手。

（1）现阶段关于增量消歧的相关研究较少。从本次文献调研结果看，增量消歧相关著述仅 30 余篇，现有研究大多面向“存量”数据的“冷启动”大批量消歧场景展开，针对增量数据的消歧研究数量相对较少。鉴于作者唯一标识并未成为作者信息的必备字段，作者同名的现象还将持续发生，作者同名增量消歧依然是当前需要解决的问题。

（2）跨语种消歧的相关研究较少。当前研究基本以单语种为主，但是随着中国国际化程度的加深，越来越多的论文在国外获得发表，跨语种消歧相关研究可能是未来的研究方向之一。

参考文献：

[1] The STM report: An overview of scientific and scholarly publishing[R/OL]. [2020-09-01]. https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf.

[2] BAGGA A, BALDWIN B. Entity-based cross-document conferencing using the vector space model[C]. Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, 1998: 79-85.

[3] SANYAL D K, BHOWMICK P K, DAS P P. A review of author name disambiguation techniques for the pub med bibliographic database[J]. Journal of information science, 2019(3): 1-28.

[4] 单嵩岩, 吴振新. 面向作者消歧和合作预测领域的作者相似度算法述评[J]. 东北师大学报(自然科学版), 2019, 51(2): 71-80.

SHAN S Y, WU Z X. Review on the author similarity algorithm in the field of author name disambiguation and research collaboration prediction[J]. Journal of northeast normal university(natural science edition), 2019, 51(2): 71-80.

[5] DELGADO A D, MONTALVO S, MARTINEZ-UNANUE R, et al. A survey of person name disambiguation on the web[J]. IEEE access, 2018, 6: 59496-59514.

[6] 沈喆, 王毅, 姚毅凡, 等. 面向学术文献的作者名消歧方法研究综述[J]. 数据分析与知识发现, 2020, 4(8) :15-27.

SHEN Z, WANG Y, YAO Y F, et al. Author name disambiguation techniques for academic literature: A review[J]. Data analysis and knowledge discovery, 2020, 4(8): 15-27.

[7] LI S, GAO C, MIAO C. Author name disambiguation using a graph model with node splitting and merging based on bibliographic information[J]. Scientometrics, 2014, 100: 15-50.

[8] CHEN Y, MARTIN J. Towards robust unsupervised personal name disambiguation[C]. Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language leaning(EMNLP-Co NLL), 2007: 190-198.

[9] 章顺瑞, 游宏梁. 基于层次聚类算法的中文人名消歧[J]. 现代图书情报技术, 2010(11): 64-68.

ZHANG S R, YOU H L. Chinese people name disambiguation by hierarchical clustering[J]. New technology of library and information service, 2010(11): 64-68.

[10] 丁海波, 肖桐, 朱靖波. 基于多阶段的中文人名消歧聚类技术的研究[C]. 第六届全国信息检索学术会(CCIR2010), 2010: 316-324.

DING H B, XIAO T, ZHU J B. A multi-stage clustering approach to Chinese person name disambiguation[C]. The 26th China conference on information retrieval(CCIR2010), 2010: 316-324.

[11] WANG X, LIU Y, WANG X, et al. Adaptive resonance theory based two-stage Chinese name disambiguation[J]. International journal, 2012(2): 83-88.

[12] LONG C, SHI L. Web person name disambiguation by relevance weighting of extended feature sets[C]. CLEF(notebook Papers/LABs/Workshops), 2010: 1-13.

[13] FERREIRA A, VELOSO A, GONCALVES M A, et al. Self-training author name disambiguation for information scarce scenarios[J].

Journal of the association for information science and technology, 2014, 65(6): 1257–1278.

[14] PEDERSEN T, PURANDARE A, KULKARNI A. Name discrimination by clustering similar contexts[C]. Computational linguistics & intelligent text processing, international conference, CICLing 2005, Mexico City, Mexico, 2005.

[15] TRAN H N, HUYNH T, DO T. Author name disambiguation by using deep neural network[C]. Asian conference on intelligent information and database systems, Springer, Cham, 2014: 123–132.

[16] 阮光册, 涂世文, 田欣, 等. 多特征融合的英文科技文献增量式人名消歧应用研究[J]. 情报杂志, 2021, 40(9): 147–153.

RUAN G C, TU S W, TIAN X, et al. Application research of incremental person name disambiguation in English scientific and technological literature based on multi feature fusion[J]. Journal of intelligence, 2021, 40(9): 147–153.

[17] 马莹莹, 吴幼龙, 唐华. 基于特征编码和图嵌入的姓名消歧方法[J]. 中国科学院大学学报, DOI:10 7523/J UCAS 2020 0019.

MA Y Y, WU Y L, TANG H. Name disambiguation based on encoding attributes and graph topology[J]. Journal of university of Chinese academy of sciences, DOI:10.7523/j. ucas. 2020.0019.

[18] MALIN B. Unsupervised name disambiguation via social network similarity[C]. Workshop on link analysis, counterterrorism, and security in conjunction with the SIAM intemational conference on data mining, 2005: 93–102.

[19] 郎君, 秦兵, 宋巍,等. 基于社会网络的人名检索结果重名消解[J]. 计算机学报,2009,32(7):1365–1373.

LANG J, QIN B, SONG W, et al. Person name disambiguation of searching results using social network[J]. Chinese journal of computers, 2009, 32(7): 1365–1373.

[20] Yao Y. Name Disambiguation method based on attribute match and link analysis[J]. Journal of software engineering and applications, 2012, 5(1): 29–32.

[21] NADIMI M H, MOSAKHANI M. A more accurate clustering method by using co–author social networks for author name disambiguation[J]. Journal of computing and security, 2015, 1(4): 102–111.

[22] PEROZZI B, AL–RFOU R, SKIENA S. Deep walk: Online learning of social representations[C]. Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, New York, USA, 2014: 701–710.

[23] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]. Proceedings of the 26th international conference on neural information processing systems, Lake Tahoe, Nevada, USA, 2013: 3111–3119.

[24] GROVER A, LESKOVEC J. Node2vec: Scalable feature learning for networks[C]. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, California, USA, 2016: 855–864.

[25] 陈丽, 朱裴松, 钱铁云, 等 基于边采样的网络表示学习模型[J]. 软件学报, 2018, 29(3): 756–771.

CHEN L, ZHU P S, QIAN T Y, et al. Edge sampling based network embedding model[J]. Journal of software, 2018, 29(3): 756–771.

[26] WANG D, PENG C, ZHU W. Structural deep network embedding[C]. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, California, USA, 2016: 1225–1234.

[27] 刘正铭, 马宏, 刘树新, 等. 一种融合节点文本属性信息的网络表示学习算法[J]. 计算机工程, 2018, 44(11): 165–171.

LIU Z M, MA H, LIU S X, et a1. A network representation learning algorithm fusing with textual attribute information of nodes[J]. Computer engineering, 2018, 44(11): 165–171.

[28] 杨欣欣, 李培峰, 朱巧明. 基于网页文本依存特征的人名消歧[J]. 计算机工程, 2012, 38(19): 133–136.

YANG X X, LI P F, ZHU Q M. Name disambiguation based on dependency feature in web page text[J]. Computer engineering, 2012, 38(19): 133–136.

[29] HAN X, ZHAO J. Web personal name disambiguation based on reference entity tables mined from the web[C]. Proceeding of the eleventh international workshop on web information and data management, 2009: 75–82.

[30] VU Q M, TAKASU A, ADACHI J. Improving the performance of personal name disambiguation using web directories[J]. Information processing and management, 2008, 44(4): 1546–1561.

[31] SHEN W, WANG J, LUO P, et al. Linking named entities with knowledge base via semantic knowledge[C]. Proceedings of the 21st

international conference on the world wide web, 2012: 449-458.

[32] 宁博, 张菲菲. 基于异构知识库的命名实体消歧[J]. 西安邮电大学学报, 2014, 19(4): 2095-6533.

NING B, ZHANG F F. Named entity disambiguation based on heterogeneous knowledge base[J]. Journal of Xi'an university of posts and telecommunications, 2014, 19(4): 2095-6533.

[33] HAN W, LIU G, MAO Y Z, et al. Attribute based Chinese named entity recognition and disambiguation[C]. The 2nd CIPS-SIGHAN joint conference on Chinese language processing, 2012: 127-131.

[34] PENG Z H, SUN L, HAN X P. A Chinese named entity recognition and disambiguation system using a two-stage method[C]. The 2nd CIPS-SIGHAN joint conference on Chinese language processing, 2012: 115-120.

[35] HE Z, LIU S, LI M, et al. Learning entity representation for entity disambiguation[C]. Proceedings of the 51st annual meeting of the association for computational linguistics, 2013: 30-34.

[36] SUN Y, LIN L, TANG D, et al. Modeling mention, context and entity with neural networks for entity disambiguation[C]. Proceedings of the 24th international conference on artificial intelligence, 2015: 1333-1335.

[37] FRANCIS L M, DURRETT G, KLEIN D. Capturing semantic similarity for entity linking with convolutional neural networks[C]. Proceedings of NAACLHLT, San Diego, 2016: 1256-1261.

[38] CHEN S, WANG J, JIANG F, et al. Improving entity linking by modeling latent entity type information[C]. Proceedings of the AAAI conference on artificial intelligence, New York, 2020: 7529-7537.

[39] GUPTA N, SINGH S, ROTH D. Entity linking via joint encoding of types, descriptions, and context[C]. Proceedings of the 2017 conference on empirical methods in natural language processing, Copenhagen, 2017: 2681-2690.

[40] LE P, TITOV I. Improving entity linking by modeling latent relations between mentions[C]. Proceedings of the 56th annual meeting of the association for computational linguistics, Melbourne, 2018: 1595-1604.

[41] GUO Z, BARBOSA D. Robust named entity disambiguation with random walks[J]. Semantic web, 2018, 9(4): 459-479.

[42] YANG X, GU X, LIN S, et al. Learning dynamic context augmentation for global entity linking[C]. Proceedings of the 2019 conference on empirical methods in natural language, Hong Kong, 2019: 271-281.

[43] LE P, TITOV I. Boosting entity linking performance by leveraging unlabeled documents[C]. Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, 2019: 1935-1945.

[44] LE P, TITOV I. Distant learning for entity linking with automatic noise detection[C]. Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, 2019: 4081-4090.

[45] 吴柯烨, 闵超, 孙建军, 等. 面向特定科研任务的著者姓名消歧方法[J]. 情报学报, 2021, 40(7): 734-744.

WU K Y, MIN C, SUN J J, et al. Method for author name disambiguation in specific research tasks[J]. Journal of the China society for scientific and technical information, 2021, 40(7): 734-744.

[46] 王若琳, 牛振东, 蔺奇卡, 等. 基于异质信息嵌入与 RNN 聚类参数预测的作者姓名消歧方法[J]. 数据分析与知识发现, 2021, 5(8): 13-24.

WANG R L, NIU Z D, LIN Q K, et al. Disambiguating author names with embedding heterogeneous information and attentive RNN clustering parameters[J]. Data analysis and knowledge discovery, 2021, 5(8): 13-24.

[47] 郭晨亮, 林欣, 殷珖. 基于异构网络的无监督作者名称消歧[J]. 华东师范大学学报(自然科学版), 2021(6): 147-160.

GUO C L, LIN X, YIN Y. Unsupervised author name disambiguation based on heterogeneous networks[J]. Journal of east China normal university(natural science), 2021(6): 147-160.

A Survey of Author Name Disambiguation Techniques of Academic Papers

WANG Xin, LU Yao, YUAN Xue, ZHAO Wanjing, CHEN Li, LIU Minjuan*

(Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081)

Abstract: [Purpose/Significance] This paper investigates the research on author name disambiguation published in recent years, and reviews the development context of relevant research from the perspective of the impact of data on author name disambiguation methods, so as to provide reference for further research. [Method/Process] The papers related to author name disambiguation were collected from English research databases such as Web of Science, Scopus, Google Academic, ACM Digital Library, IEEE Xplore, ScienceDirect, Scopus and Springer Link, and Chinese research databases such as CNKI, CQVIP and WANFANG. The search results cover the relevant papers published from 1998 to 2021. On the premise of giving consideration to authority, influence and novelty, 46 publicationswere selected for review. There are many types and structures of author name disambiguation data. For example, literature feature information is generally presented in unstructured text, and the extracted features can be stored and represented in two-dimensional tables; Citation information and interpersonal relationship are network relational data, which can be stored and represented by graphs, key value pairs or two-dimensional tables. The fundamental reason for different data structures lies in their semantic differences, but the data structure itself determines its applicable algorithm. According to the structure of characteristic data used in the author name disambiguation task and the different corresponding data processing algorithms, the relevant research is divided into three categories: 1) disambiguation method based on literature characteristics, 2) disambiguation method based on social network and 3) disambiguation method by integrating external knowledge. The impact of data on the author name disambiguation method is examined from the data level. [Results/Conclusions] The analysis found that with the progress of technology, deep learning methods have been widely used. Compared with the improvement of the model, the feature learning and representation based on deep learning can significantly improve the effect of the author name disambiguation algorithm. In addition, in order to overcome the problem of insufficient data utilization by a single method and improve the utilization efficiency of data, the three methods show the trend of mutual combination and complementary gain. From the literature research results, there are few related studies on incremental author name disambiguation and multi-language author name disambiguation, which could be one of the directions for further research.

Keywords: knowledge organization; author name disambiguation; person name disambiguation